W340–W345 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh447

ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry

Richard D. LeDuc, Gregory K. Taylor[1], Yong-Bin Kim[1], Thomas E. Januszyk[1], Lee H. Bynum[1], Joseph V. Sola[1], John S. Garavelli[2] and Neil L. Kelleher[1,*]

W.M. Keck Center for Comparative and Functional Genomics and [1]Department of Chemistry, University of Illinois, Urbana, IL, 61801, USA and [2]EMBL Outstation–EBI, Hinxton, Cambridge CB10 1SD, UK

Received February 16, 2004; Revised and Accepted April 20, 2004

ABSTRACT

ProSight PTM (https://prosightptm.scs.uiuc.edu/) is a web application for identification and characterization of proteins using mass spectra data from 'top-down' fragmentation of intact protein ions (i.e. without any tryptic digestion). ProSight PTM has many tools and graphical features to facilitate analysis of single proteins, proteins in mixtures and proteins fragmented in parallel. Sequence databases from across the phylogenetic tree are supported, with a new database strategy of 'shotgun annotation' used to assist characterization of wild-type proteins. During a database search, data from divergent sources regarding potential mass differences such as polymorphisms, alternate splicing and post-translational modifications are utilized. The user can optionally control how much of this biological variability should be searched.

INTRODUCTION

The identification and characterization of intact proteins from tandem mass spectrometry experiments (MS/MS) in top-down proteomics is a rapidly emerging alternative to the traditional bottom-up techniques. In bottom-up proteomics, intact proteins are first digested with tryptic enzymes, and the resulting mixtures of 1–3 kDa peptides are analyzed using various combinations of chromatography followed by MS/MS analysis. Each peptide MS/MS spectrum is then compared to the predicted fragment ions of tryptic peptides from proteins in a sequence database. Identification is accomplished by locating the protein in the database that best matches the observed tryptic peptides or their fragment ions. Peptides with anomalous masses can be used to identify post-translational modifications (PTMs) (1,2). Unfortunately, most bottom-up techniques typically examine only between 5 and 70% of the possible locations of PTMs (3).

Top-down approaches typically deliver 100% sequence coverage for proteins <70 kDa and most often utilize Fourier transform MS. Top-down proteomics is based on tandem MS (MS/MS) where a mixture of intact proteins is first separated in the gas phase, and then specific protein ions of known intact mass are isolated and fragmented. By comparing the intact mass and the resulting fragmentation data to a database of known sequences, not only can the protein be identified, but coding polymorphisms (cSNPs) and PTMs on the protein can be simultaneously characterized.

ProSight PTM (https://prosightptm.scs.uiuc.edu/) is the first web application that combines search engines and a browser environment to allow researchers to analyze data from top-down analysis of >10 kDa proteins. ProSight PTM is centered on organism-specific databases of protein- and site-specific information. This organism-specific information is derived from known and predicted processing events such as coding polymorphisms, alternate splicing and of course PTMs. An example identification of human 40S ribosomal protein S28 will be used to illustrate these tools.

IMPLEMENTATION

ProSight PTM is composed of a number of independent search and characterization applications that are bound together by a Perl CGI interface. Additionally, several utility applications, designed to assist in manipulating MS data, are also provided. Figure 1 shows the organization of major ProSight PTM components.

The ProSight Warehouse

ProSight Warehouse (4) incorporates diverse information necessary for identification and characterization of typical

*To whom correspondence should be addressed. Tel: +1 217 244 3927; Fax: +1 217 244 8068; Email: kelleher@scs.uiuc.edu

wild-type proteins. This includes protein sequences, average and monoisotopic masses, and known or putative PTMs. The data on each organism in ProSight Warehouse are in a dedicated MySQL database. MySQL (www.mysql.com) was chosen because it is open source, free and has many supporting tools and application programming interfaces (APIs). Warehouse uses version 11.18 distribution 3.23.52 of MySQL for Linux.

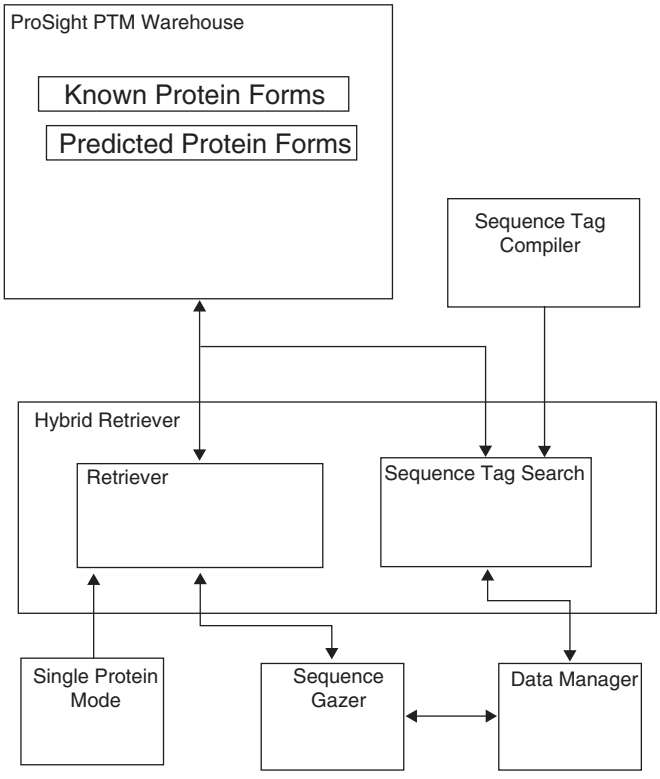

Figure 1. Organization of the major ProSight PTM components. ProSight PTM contains six major components. ProSight Warehouse is a data warehouse of known and predicted protein forms of many different organisms. The warehouse is queried via either Retriever or the Sequence Tag Search. Retriever performs absolute mass searches by comparing observed against predicted fragment masses derived from a set of candidate protein forms selected using knowledge of the intact mass. Retriever interacts with Single Protein Mode to allow the iterative testing of hypotheses to explain mass discrepancies, and thus characterize an identified protein. The Sequence Tag Search identifies candidate proteins based on all possible sequence tags consistent with the fragmentation data. The Data Manager organizes data within the system and maintains data continuity between user sessions.

Sequence data for a given organism can be stored with either a simple or highly annotated schema. The simple schema format uses a two-tiered approach that assumes that many different protein forms can be generated from a single gene. This approach is based on genes, each with one basic sequence derived from genomic annotation, having many different protein forms from known or putative PTMs. However, this does not fully reflect the nature of eukaryotic proteins since a single gene could have more than one basic sequence due to alternate splicing, conflicts, sequence variants, etc. This is resolved with our highly annotated three-tiered model. In this model, one gene can have several basic sequences and each basic sequence may have many different protein forms.

Simple schemas have been developed for yeast (*Saccharomyces cerevisiae*, SGD), *Escherichia coli* (Swiss-Prot), *E. coli* (Wisconsin), *Arabidopsis thaliana* (Swiss-Prot), *Bacillus subtilis*, *Methanococcus jannaschii*, *Mycoplasma pneumoniae*, *Shewanella oneidensis* (Swiss-Prot) and *Methanosarcina acetivorans* (MIT Center for Genome Research). The yeast database includes an additional 52 experimentally verified phosphorylation cases (5), the experimental cases from Meng *et al*. (6) and the signal peptide forms that are predicted by Sigcleave (http://bioweb.pasteur.fr/seqanal/interfaces/sigcleave.html). All other organisms with simple schemas include basic sequences initial methionine cleavage and N-terminal acetylation.

Highly annotated schemas, such as human and mouse, are based on the three-tier model and were developed as a response to the large number of factors that can change the mass of a protein. These databases allow for the following: phosphorylation, acetylation, methylation, conversion to a pyruvoyl group, hypusine, formylation, diphthamide, lipid (myristate, palmitate, farnesyl, geranyl–geranyl, GPI-anchor), selenocysteine, initial methionine cleavage and signal peptides. All of the PTMs are handled using the PTM identifiers within the RESID database (7).

Table 1 shows the current number of protein forms and the size of each database.

ProSight Retriever

ProSight Retriever (4) was designed using object-oriented principles and implemented in C++ using the GNU gcc compiler for a Linux server. Retriever connects to ProSight Warehouse using ODBC.

Table 1. Current number of protein forms and size of databases stored in the ProSight Warehouse

Organism	Entries	Size (MB)	Data source	Remark
Saccharomyces cerevisiae	31 050	16.7	SGD	Simple
Shewanella oneidensis	12 504	4.5	Swiss-Prot/TrEMBL	Simple
Methanosarcina acetivorans	11 701	4.6	MIT Center for Genomic Research	Simple
Methanococcus jannaschii	4638	1.8	Olsen	Simple
Escherichia coli	22 621	8.0	Swiss-Prot	Simple
Escherichia coli	22 301	7.6	Wisconsin	Simple
Mycoplasma pneumoniae	1758	0.9	Herrmann	Simple
Bacillus subtilis	10 301	3.8	Kunst/Danchin	Simple
Arabidopsis thaliana	72 705	34.7	Swiss-Prot/TrEMBL	Simple
Homo sapiens	134 180	84.6	Swiss-Prot/TrEMBL	Heavily annotated
Homo sapiens	116 279	74.5	HPI	Heavily annotated
Mus musculus	82 435	44.0	Swiss-Prot/TrEMBL	Heavily annotated

Retriever is used for absolute mass searching. It first isolates a set of candidate protein-form sequences from the ProSight Warehouse that are within a user-specified tolerance of the observed intact mass (e.g. ±2000 Da). To determine which sequence best accounts for the observed data, all theoretical *b*/*y* or *c*/*z*• fragment ions are generated for each candidate sequence. The scoring process counts the number of observed fragment masses that match theoretical fragment masses within a user-specified tolerance (e.g. 25 parts-per-million, ppm.). In cases where an observed mass matches multiple theoretical fragments, each case is output, but only one match is counted when computing the score. The resulting *P*-score, derived from a Poisson-based statistical model and representing the probability that a random sequence could account for the observed data, is described elsewhere (8). These fragment ions are created from either 'threshold' methods for gas phase protein fragmentation (*b*- and *y*-type fragment ions) or electron capture dissociation (ECD) (*c*- and *z*•-type), a relatively new method for MS/MS (9).

In the case of the S28 ribosomal protein, 148 fragment ion masses were observed after fragmentation by ECD. The intact mass of 7949.42 Da (monoisotopic mass) was used to search the highly annotated human database from Swiss-Prot/TrEMBL with a mass window of ±5000 Da. The error tolerance on the fragment ions was 30 ppm. Figure 2 shows the output from this search. The best match for this data was the S28 ribosomal protein with an acetylation on the N-terminal methionine.

Sequence Tag Search

Sequence tag ladders are formed from gas phase cleavage of adjacent backbone bonds and allow short amino acid sequences to be determined from fragment ion MS/MS data (10). They consist of one or more elements, each containing either amino acid sequences or a mass value of a fragment difference that is consistent with two or more amino acids (called here 'missing rung' values). Sequence tags, in ProSight PTM, have the restriction that they cannot be composed completely of missing rung values.

Sequence Tag Search operates in two steps: sequence tag compilation, followed by a database search for the resulting tags. In the compilation step, sequence tags are predicted from the observed fragment ions. During the search, all sequences consistent with the fragment ions are identified.

The compilation process consists of building a graph where the observed fragment masses are nodes and amino acids represent edges. In this algorithm, the fragment ions are read from a .pkl file and then sorted by mass, and the set of all single and double amino acid sequences is read from a flat text file. When comparing two vertices, the mass difference is then used to search the set of potential edges. The tolerance of the search is relative (in ppm.) to the larger of the two masses being compared. If the difference of two fragments is consistent with one or more single or double amino acids then an edge connecting the two fragment vertices is added to the graph. This process of adding edges results in a multigraph where each path corresponds to a possible sequence tag that is consistent with the observed data. Because there is a user-specified tolerance when comparing fragment mass differences, leucine and isoleucine are not the only two amino acids that might be considered isomeric. Any set of amino acids that cannot be separated under the current tolerance is reported as a character class, and is noted within square brackets with a pipe separating isomers, e.g. [K | Q, both nominally 128 Da].

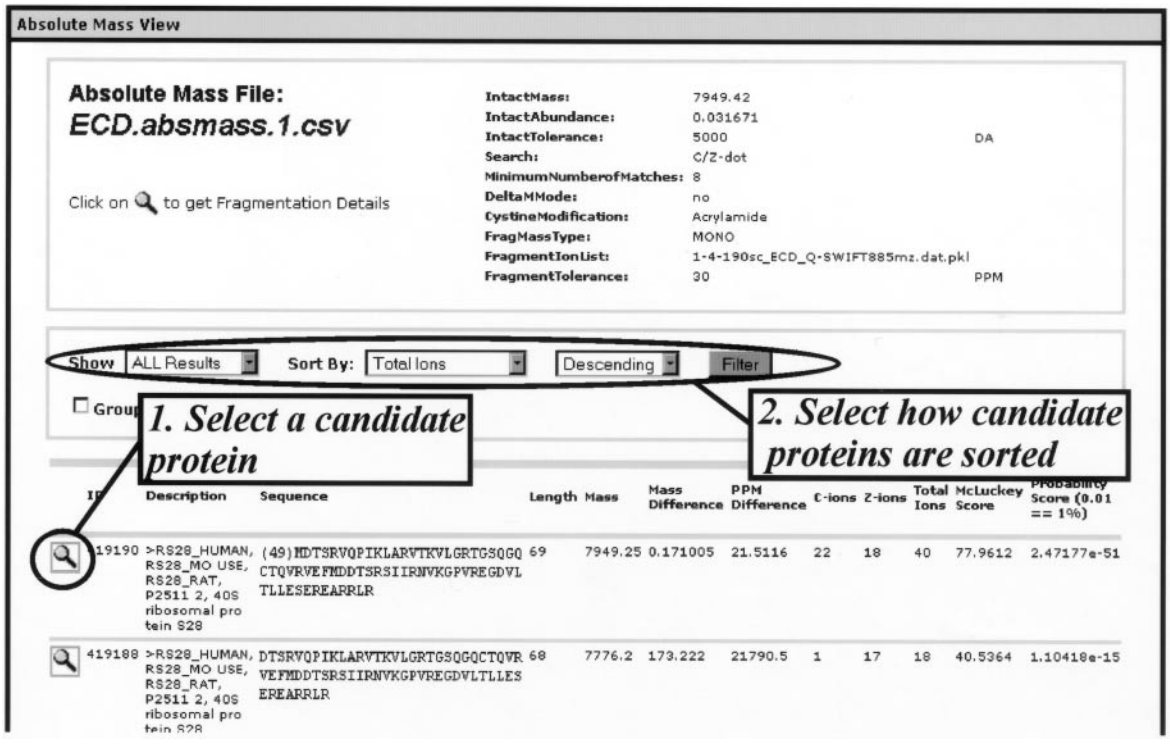

Figure 2. Results of an absolute mass search on the S2B ribosomal protein. Label 1 identifies the link icon used to pick results of interest for further display. Label 2 identifies the fields that determine how the retrieved proteins are sorted.

After all of the fragment pairs have been compared, the tags are read using a depth-first search. For simplicity, tags that do not contain a single amino acid or have two or fewer rungs are not output.

Once a list of one or more sequence tags has been created, the tags can be searched against any ProSight Warehouse database. In the search, each basic sequence in the database is queried against the entire list of tags and their reverse in turn. For each tag, within each sequence, the first element containing an amino acid is located, and two lists are created. The first is the list of missing rung values to the left of the first amino acid element, and the second is the remainder of the tag. For each occurrence of the first amino acid element in the sequence, the left-hand amino acid masses are checked for consistency with the missing rungs to the left of the hit. If the sequence and the missing rung values match, then the right side of the tag is compared to the sequences to the right of the hit. If the sequence is consistent with the entire tag, then the score of the portion of the sequence matching the tag is calculated and added to the sequence's total score. In this way, the more tags that match the sequence, the higher the sequence's final score.

To score sequence tags, each amino acid is given a probability based on prior knowledge (11). The sequence of amino acids found to be consistent with the tag is then scored as:

$$\text{tag} = -\ln\left(\prod^{n} p_i\right)\left(\frac{n_i}{l}\right),$$

where p_i is the probability associated with the i-th amino acid, l is the overall length of the sequence and n_i is the length of sequences in the tag. Since multiple tags may hit the sequence, each tag should be weighted by the number of independent possibilities for the tag to hit the sequence, which is approximated by l/n. The final score for a query is then the sum of all tag scores that hit the sequence.

Hybrid Retriever

Hybrid Retriever performs a sequence tag search to retrieve unmodified candidate proteins from the database. After these candidates have been identified, Retriever is used to read all of the modified protein forms generated from the unmodified candidate proteins from the database. The traditional absolute mass search is then performed on these modified sequences.

Data Manager

In addition to providing analysis tools, ProSight PTM also provides data management and visualization tools. Data Manager allows users to maintain ongoing research on the ProSight PTM server. Files can be stored and organized to the researcher's preference. Search results can be kept for comparison, or further analysis, and data can be up- and downloaded from the server, all through an intuitive graphical interface. Most of the data management and visualization tools are contained within the Perl interface. Graphical output is generated using the Perl GD module (http://stein.cshl.org/WWW/software/GD/), and all images are exported as .png files (http://www.libpng.org/pub/png/).

Utility functions

ProSight PTM offers a number of utility applications designed to assist in analyzing top-down proteomics data. Each of these tools is a separate application that is called by a convenient web-form CGI interface.

PKL and PRL Makers. Intact masses and fragmentation lists from non-MIDAS (12) systems can easily be saved as PRL- and PKL-formatted files using these two tools.

Noise Reducer. Noise Reducer, written in Ocaml, removes duplicate masses (masses within a user-specified tolerance) from a .pkl file. Further, it removes water and ammonia losses typically observed in 'threshold' MS/MS spectra. The program also detects when two masses differ by 1 Da (a common error in automated interpretation of MS/MS spectra with fragment ions >10 kDa).

Ion Predictor. Takes a sequence and predicts all possible *b*/*y* and or *c*/*z*• ions that could be generated. It reports either average or monoisotopic masses.

Single Protein Mode. This utility takes intact mass data, fragment data and an amino acid sequence and matches the data to the sequence. Further, hypothetical PTMs can be placed on the sequence and then tested against the observed data. In this way, an identified protein can be completely characterized in an interactive, user-driven environment. Single Protein Mode can be used to test for phosphorylation, methylation, dimethylation, hexose or any desired 'custom' mass shift. Single Protein Mode uses the Perl web interface to send specialized queries to Retriever and uses Perl GD heavily for the graphical placement of hypothetical PTMs.

OPERATION

There are many ways to use ProSight PTM to identify and characterize proteins. Once the user has completed the authenticated login to the system, the Data Manager is presented. From this screen, the user can create files and directories or move, copy, delete or compress data, as well as up- and download files. On the left side of the screen is a list of tools available to the user. Input data may be either uploaded from MIDAS .pkl and .prl files (12) or manually entered. If desired, the PRL and PKL Makers can be used on manually entered data to save them as the appropriate file type. The Data Manager screen displays all files stored in the current working folder, as well as navigational information to other user-created folders. Only the files displayed in the current folder will be available to the tools selected. Each tool will automatically provide the user with a list of appropriate files from the current directory.

Conducting an absolute mass search

Some semi-automated top-down experiments result in sparse fragmentation data. On a 20 kDa protein, between 0 and >200 fragment ions can be produced, corresponding to 0 to >100 distinct masses. We recommend first using the absolute mass search. Frequently, this alone will identify the correct protein even for modified mammalian proteins. An identification is considered legitimate if a *P*-score is reported to be <0.01 (i.e. there is only a 1% chance that the observed protein matches the data by chance). The absolute mass search often fails due to terminal modifications/sequence events at the termini or if the observed mass is wildly different from the mass in the database (e.g. alternative splicing and proteolysis). In such cases,

Sequence Tag Search offers an alternative method of identification.

Figure 2 shows a portion of the output file from the human database search with the data generated from the S28 ribosomal protein. Each row represents information from a single protein found by Retriever. Notice that the results can be sorted by any of the numeric result fields. In this case, the best match (an N-terminally acetylated form of ribosomal protein S28) had 40 total fragment ion hits at 30 p.p.m. mass accuracy. The resulting *P*-score was quite good: 2.47×10^{-51}. The next highest match was also protein S28, but without the start methionine or any acetylation. The *P*-score for this match was 1.10×10^{-15}—from 18 total fragment ion matches—well above the score for the best match. Any record can be selected and viewed in detail in the Graphical Fragment Mapper (Figure 3). Here, some of the

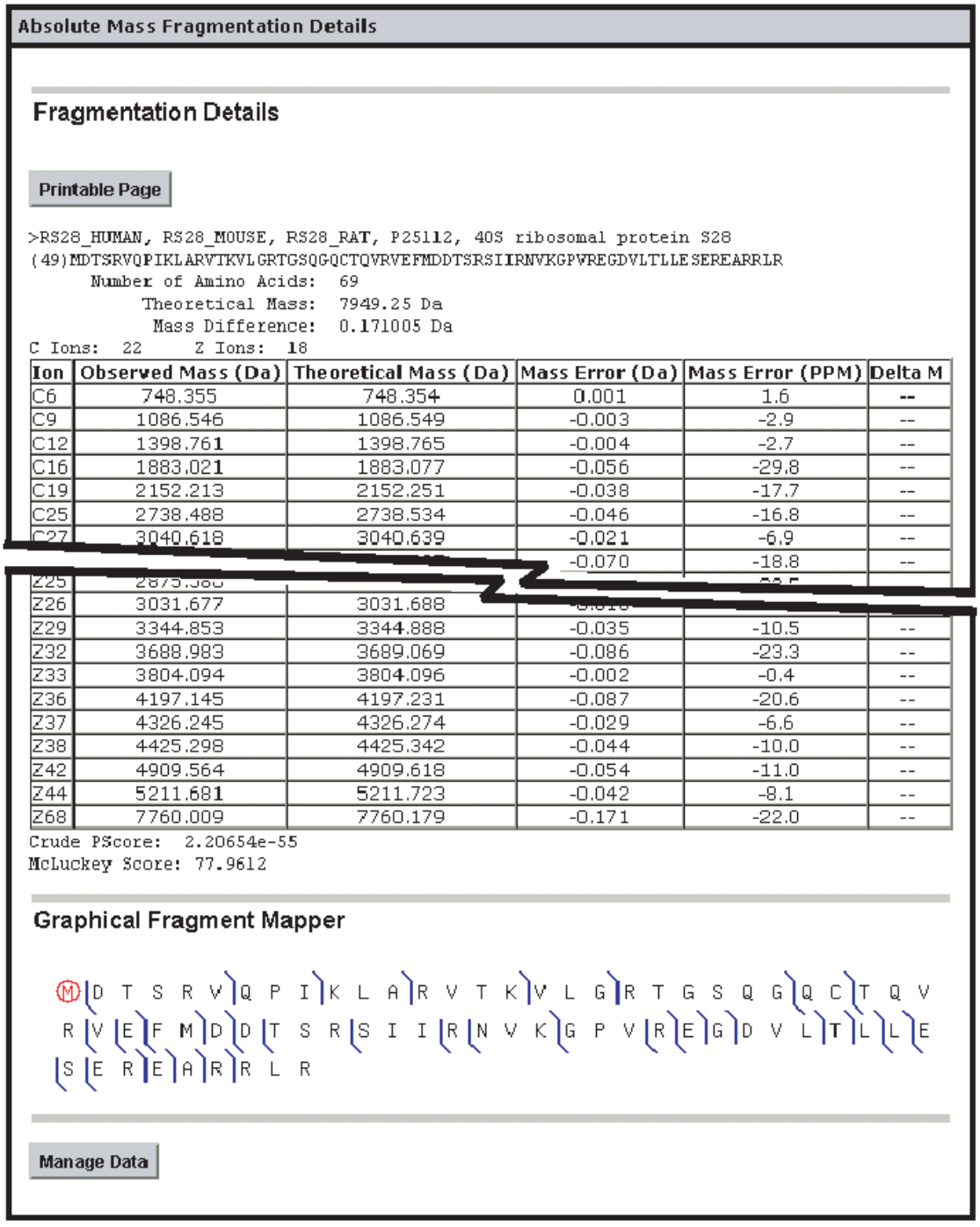

Figure 3. Fragmentation details. For each absolute mass search result, basic information about the protein is displayed. Additionally, a table listing the observed and theoretical fragments and a visual representation are also provided. The circled red M signifies a post-translational modification at sequence position 1. Drilling down on the image of M will link the user to the RESID (13) database for further information.

40 matching fragment ions are shown with the accompanying graphical fragment map.

Conducting a Sequence Tag Search

The first step in identifying a sequence using sequence tags is to compile sequence tags from the observed fragment ions. When compiling sequence tags, use a tolerance from 1 to 25 p.p.m. for high-accuracy MS instruments (e.g. FT–ICR MS). With a large number of fragment masses, a tight intact tolerance is critical. Too high a tolerance will result in a large number of spurious tags; too small a tolerance will miss valid tags. If a large number of tags is returned, specifying a charge tolerance will help eliminate spurious tags (4). In top-down MS/MS, charge tends to increase slowly with mass (on average 1 charge with every 1 kDa). The charge tolerance specifies the maximum number of charges that a peptide may gain between two fragments. Use a charge tolerance of 1 to reduce the number of spurious tags generated.

Once tags have been compiled, use Sequence Tag Search to query the ProSight Warehouse. The search engine will handle any number of tags resulting from the same set of observed fragment ions. If there are only a few tags, consider setting the minimum score feature to a low value such as 0.001. If the set of tags contains one or more short tags of two or three amino acids, many spurious hits may result. Increasing the minimum score will remove the spurious hits.

Tolerance is important when comparing missing rung masses i.e. a 2–4 residue gap in a sequencing ladder to Warehouse sequences. To be a match, the sum of the sequence amino acid masses must be plus or minus the tolerance of the missing rung mass. Consider starting with a tolerance of 0.3 Da. Increasing the tolerance will allow more sequences to fall within the missing rung mass, while decreasing it will restrict this number.

Finishing

Once a protein has been identified, any remaining mass discrepancies can be explored using Single Protein Mode. By iteratively testing PTM hypotheses against the observed data, the complete protein can be characterized.

OBTAINING AN ACCOUNT

Accounts for academic users are provided free of charge. Email prosightptm@scs.uiuc.edu for account information.

FUTURE DIRECTIONS

An effort is being made to alter the ProSight PTM interface to better meet the needs of the user community. This will require more sophisticated graphical capabilities, improved interconnection of the ProSight PTM tools, automatic linking of results screens to relevant external data sources and the streamlining of steps within an analysis.

Further, one area of current foucs is expanding ProSight Warehouse to house even more sources of biological and informatic variability. This will require the development of smarter algorithms to generate and search a larger number of biologically possible explanations of mass discrepancies between wild-type proteins and predicted mass values from fully sequenced genomes.

After a protein form has been identified, discrepancies often arise between the identified sequence and the observed data. Currently the Hybrid Retriever algorithm gives a superabundance of bioinformatics predictions. An altered version of the existing algorithm, in development, will determine which events (SNPs, PTMs, etc.) most likely give rise to the observed data. This idea extends the PROCLAME (http://proclame.unc.edu/) philosophy by predicting gene/location-specific modifications and evaluating a prediction set by the resulting degree of sequence coverage.

ACKNOWLEDGEMENTS

We thank these sources of support: the Searle Scholars Program Foundation, the Burroughs Wellcome Fund, the University of Illinois, the Packard Foundation and the National Institutes of Health (GM 067193).

REFERENCES

1. Washburn,M.P. and Yates,J.R. (2000) Analysis of the microbial proteome. *Curr. Opin. Microbiol.*, **3**, 292–297.
2. Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.
3. Fountoulakis,M., Takacs,M.F., Berndt,P., Langen,H. and Takacs,B. (1999) Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis*, **20**, 2181–2195.
4. Taylor,G.K., Kim,Y.-B., Forbes,A.J., Meng,F., McCarthy,R. and Kelleher,N.L. (2003) Web and database software for identification of intact proteins using 'top-down' mass spectrometry. *Anal. Chem.*, **75**, 4081–4086.
5. Ficarro,S.B., McCleland,M.L., Stukenberg,P.T., Burke,D.J., Ross,M.M., Shabanowitz,J., Hunt,D.F. and White,F.M. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **20**, 301–305.
6. Meng,F., Cargile,B.J., Patrie,S.M., Johnson,J.R., McLoughlin,S.M. and Kelleher,N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal. Chem.*, **74**, 2923–2929.
7. Garavelli,J.S., Hou,Z., Pattabiraman,N. and Stephens,R.M. (2001) The RESID database of protein structure modification and the NRL-3D Sequence–Structure database. *Nucleic Acids Res.*, **29**, 199–201.
8. Meng,F., Cargile,B.J., Miller,L.M., Forbes,A.J., Johnson,J.R. and Kelleher,N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.*, **19**, 952–957.
9. Zubarev,R.A., Kelleher,N.L. and McLafferty,F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc., (Communication)*, **120**, 3265–3266.
10. Mann,M. and Wilm,M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, **66**, 4390–4399.
11. McCaldon,P. and Argos,P. (1988) Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins*, **4**, 99–122.
12. Senko,M.W., Canterbury,J.D. and Guan,S. (1996) A high-performance modular data system for Fourier transform ion cyclotron resonance mass spectrometry. *Rapid Commun. Mass Spectrom.*, **10**, 1839–1844.
13. Garavelli,J.S. (2003) The RESID database of protein modifications: 2003 developments. *Nucleic Acids. Res.*, **31**, 499–501.